SUPPLEMENTARY MATERIAL FOR "STRONG ORACLE OPTIMALITY OF FOLDED CONCAVE PENALIZED ESTIMATION"

By Jianqing Fan*, Lingzhou Xue and Hui Zou†

Princeton University and University of Minnesota

In this supplementary material we give the complete proof of Theorem 5 in Fan, Xue and Zou (2013).

Theorem 5 Let $m = \max_{(i,j)} |x_{ij}|$. Under Assumption (A2) and

$$\text{(C2)}\quad \kappa_{logit} = \min_{\boldsymbol{u}\neq \boldsymbol{0}: \|\boldsymbol{u}_{\mathcal{A}^c}\|_{\ell_1} \leq 3\|\boldsymbol{u}_{\mathcal{A}}\|_{\ell_1}} \frac{\boldsymbol{u}'\nabla^2 \ell_n^{logit}(\boldsymbol{\beta})\boldsymbol{u}}{\boldsymbol{u}'\boldsymbol{u}} \in (0,\infty),$$

if $\lambda_{lasso} \leq \frac{\kappa_{logit}}{20ms}$, with probability at least $1-2p\cdot\exp(-\frac{n}{2M}\lambda_{lasso}^2)$, we have

$$\|\widehat{\boldsymbol{\beta}}^{lasso} - \boldsymbol{\beta}^\star\|_{\ell_2} \leq 5\kappa_{logit}^{-1} s^{1/2}\lambda_{lasso}.$$

<u>Proof of Theorem 5</u>. By definition, it obviously holds that

$$\ell_n(\widehat{\boldsymbol{\beta}}^{lasso}) + \lambda_{lasso}\|\widehat{\boldsymbol{\beta}}^{lasso}\|_{\ell_1} \leq \ell_n(\boldsymbol{\beta}^\star) + \lambda_{lasso}\|\boldsymbol{\beta}^\star\|_{\ell_1}.$$

Using the convexity of $\ell_n(\cdot)$, we obtain

$$(\nabla \ell_n(\boldsymbol{\beta}^\star))'(\widehat{\boldsymbol{\beta}}^{lasso} - \boldsymbol{\beta}^\star) + \lambda_{lasso}\|\widehat{\boldsymbol{\beta}}^{lasso}\|_{\ell_1} \leq \lambda_{lasso}\|\boldsymbol{\beta}^\star\|_{\ell_1}.$$

This entails that on the event

$$\{\|\frac{1}{n}\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^\star))\|_{\max} \leq \frac{1}{2}\lambda_{lasso}\} \tag{1}$$

we have

$$-\frac{1}{2}\lambda_{lasso}\|\widehat{\boldsymbol{\beta}}^{lasso} - \boldsymbol{\beta}^\star\|_{\ell_1} + \lambda_{lasso}\|\widehat{\boldsymbol{\beta}}^{lasso}\|_{\ell_1} \leq \lambda_{lasso}\|\boldsymbol{\beta}^\star\|_{\ell_1},$$

or

$$\frac{1}{2}\|\widehat{\boldsymbol{\beta}}^{lasso} - \boldsymbol{\beta}^\star\|_{\ell_1} \leq \|\boldsymbol{\beta}^\star\|_{\ell_1} - \|\widehat{\boldsymbol{\beta}}^{lasso}\|_{\ell_1} + \|\widehat{\boldsymbol{\beta}}^{lasso} - \boldsymbol{\beta}^\star\|_{\ell_1}.$$

*Supported by the National Institutes of Health grant R01-GM072611 and National Science Foundation grants DMS-1206464 and DMS-0704337.

†Supported by the National Science Foundation grant DMS-0846068 and a grant from the Office of Naval Research.

Using the fact that $|\beta^\star_j| - |\hat\beta^{lasso}_j| + |\beta^\star_j - \hat\beta^{lasso}_j| = 0$ for any $j \in \mathcal{A}^c$, we conclude that

$$\frac{1}{2}\|\widehat{\boldsymbol\beta}^{lasso} - \boldsymbol\beta^\star\|_{\ell_1} \le 2\|\widehat{\boldsymbol\beta}^{lasso}_{\mathcal{A}} - \boldsymbol\beta^\star_{\mathcal{A}}\|_{\ell_1}$$

where we denote $\widehat{\boldsymbol\beta}^{lasso} = (\widehat{\boldsymbol\beta}^{lasso}_{\mathcal{A}}, \widehat{\boldsymbol\beta}^{lasso}_{\mathcal{A}^c})$. The last inequality is equivalent to

$$\|\widehat{\boldsymbol\beta}^{lasso}_{\mathcal{A}^c}\|_{\ell_1} \le 3\|\widehat{\boldsymbol\beta}^{lasso}_{\mathcal{A}} - \boldsymbol\beta^\star_{\mathcal{A}}\|_{\ell_1}. \tag{2}$$

In what follows, our aim is to derive the upper bound

$$\|\widehat{\boldsymbol\beta}^{lasso} - \boldsymbol\beta^\star\|_{\ell_2} \le 5\kappa^{-1}_{logit} s^{1/2}\lambda_{lasso}$$

under the event (1). Then the desired probability bound can be obtained by using the Hoeffding's bound in the proof of Theorem 4 of Fan et al. (2013).

Now we consider a map $F : \mathbb{R}^p \to \mathbb{R}$ satisfying

$$F(\boldsymbol\Delta) = \ell_n(\boldsymbol\beta^\star + \boldsymbol\Delta) - \ell_n(\boldsymbol\beta^\star) + \lambda_{lasso}(\|\boldsymbol\beta^\star + \boldsymbol\Delta\|_{\ell_1} - \|\boldsymbol\beta^\star\|_{\ell_1}).$$

In addition, we define $\widehat{\boldsymbol\Delta} = \arg\min_{\boldsymbol\Delta} F(\boldsymbol\Delta)$. Then by definition we have $\widehat{\boldsymbol\Delta} = \widehat{\boldsymbol\beta}^{lasso} - \boldsymbol\beta^\star$. Since $F(\mathbf{0}) = 0$, $F(\widehat{\boldsymbol\Delta}) \le F(\mathbf{0}) = 0$. By Lemma 4 of Negahban et al. (2012), because $\|\widehat{\boldsymbol\Delta}_{\mathcal{A}^c}\|_{\ell_1} \le 3\|\widehat{\boldsymbol\Delta}_{\mathcal{A}}\|_{\ell_1}$ as in (2) and convexity of $F(\boldsymbol\Delta)$, it suffices to show that

$$F(\boldsymbol\Delta) > 0$$

for any $\boldsymbol\Delta \in \mathcal{D}$, where

$$\mathcal{D} = \{\boldsymbol\Delta \in \mathbb{R}^p : \|\boldsymbol\Delta_{\mathcal{A}^c}\|_{\ell_1} \le 3\|\boldsymbol\Delta_{\mathcal{A}}\|_{\ell_1} \text{ and } \|\boldsymbol\Delta\|_{\ell_2} = 5\kappa^{-1}_{logit} s^{1/2}\lambda_{lasso}\}.$$

To this end, we first obtain a lower bound for $\|\boldsymbol\beta^\star + \boldsymbol\Delta\|_{\ell_1} - \|\boldsymbol\beta^\star\|_{\ell_1}$, i.e.,

$$\begin{aligned} \|\boldsymbol\beta^\star + \boldsymbol\Delta\|_{\ell_1} - \|\boldsymbol\beta^\star\|_{\ell_1} &= \|\boldsymbol\beta^\star_{\mathcal{A}} + \boldsymbol\Delta_{\mathcal{A}}\|_{\ell_1} + \|\boldsymbol\Delta_{\mathcal{A}^c}\|_{\ell_1} - \|\boldsymbol\beta^\star_{\mathcal{A}}\|_{\ell_1} \\ &\ge \|\boldsymbol\Delta_{\mathcal{A}^c}\|_{\ell_1} - \|\boldsymbol\Delta_{\mathcal{A}}\|_{\ell_1} \end{aligned} \tag{3}$$

Next, we derive a lower bound for $\ell_n(\boldsymbol\beta^\star + \boldsymbol\Delta) - \ell_n(\boldsymbol\beta^\star)$. To simplify notation, we define $G(u) = \ell_n(\boldsymbol\beta^\star + u\boldsymbol\Delta)$. Recall that $\psi''(t) = \theta(t)(1-\theta(t))$ and $\psi'''(t) = \theta(t)(1-\theta(t))(2\theta(t)-1)$ with $\theta(t) = (1+\exp(t))^{-1}$. Then we have

$$\begin{aligned} G''(u) &= \frac{1}{n}\sum_i \psi''(\boldsymbol{x}_i'(\boldsymbol\beta^\star + u\boldsymbol\Delta)) \cdot (\boldsymbol{x}_i'\boldsymbol\Delta)^2 \\ G'''(u) &= \frac{1}{n}\sum_i \psi'''(\boldsymbol{x}_i'(\boldsymbol\beta^\star + u\boldsymbol\Delta)) \cdot (\boldsymbol{x}_i'\boldsymbol\Delta)^3 \end{aligned}$$

By using the simple fact that

$$0 \le |\psi'''(t)| \le \psi''(t),$$

we have

$$|G'''(u)| \le \max_i |\boldsymbol{x}_i'\boldsymbol{\Delta}| \cdot G''(u) \le m\|\boldsymbol{\Delta}\|_{\ell_1} \cdot G''(u).$$

Note that by the definition of $\mathcal{D}$,

$$\|\boldsymbol{\Delta}\|_{\ell_1} = \|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\ell_1} + \|\boldsymbol{\Delta}_{\mathcal{A}^c}\|_{\ell_1} \le 4\|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\ell_1} \le 4ms^{1/2}\|\boldsymbol{\Delta}\|_{\ell_2}.$$

Let $z = 4ms^{1/2}\|\boldsymbol{\Delta}\|_{\ell_2} = 20m\kappa_{logit}^{-1}s\lambda_{lasso} > 0$. Then we have

$$|G'''(u)| \le zG''(u)$$

By Lemma 1 of Bach (2010), for any convex three times differentiable function $g(u)$ satisfying $|g'''(u)| \le Sg''(u)$ for some $S > 0$, we have

$$g(u) - g(0) - g'(0)u \ge g''(0) \cdot S^{-2}\{\exp(-uS) + uS - 1\}.$$

Here we consider $g(u) = G(u)$ and $S = z$. Let $u = 1$, and then we obtain

$$G(1) - G(0) - G'(0) \ge G''(0) \cdot h(z), \tag{4}$$

where $h(z) = z^{-2}(\exp(-z) + z - 1)$. By simple calculation it can be shown that $h(z)$ is a decreasing function in $z > 0$. Given that $z \le 1$ holds by assumption on λ_{lasso}, we have

$$h(z) \ge h(1) = \exp(-1) > 1/3.$$

By definition $G(1) = \ell_n(\boldsymbol{\beta}^\star + \boldsymbol{\Delta})$, $G(0) = \ell_n(\boldsymbol{\beta}^\star)$, $G'(0) = (\nabla\ell_n(\boldsymbol{\beta}^\star))'\boldsymbol{\Delta}$ and $G''(0) = \boldsymbol{\Delta}'\nabla^2\ell_n(\boldsymbol{\beta}^\star)\boldsymbol{\Delta}$. Thus, we can re-write (4) as

$$\begin{aligned} \ell_n(\boldsymbol{\beta}^\star + \boldsymbol{\Delta}) - \ell_n(\boldsymbol{\beta}^\star) &\ge (\nabla\ell_n(\boldsymbol{\beta}^\star))'\boldsymbol{\Delta} + h(z)\boldsymbol{\Delta}'\nabla^2\ell_n(\boldsymbol{\beta}^\star)\boldsymbol{\Delta} \\ &> (\nabla\ell_n(\boldsymbol{\beta}^\star))'\boldsymbol{\Delta} + \frac{1}{3}\boldsymbol{\Delta}'\nabla^2\ell_n(\boldsymbol{\beta}^\star)\boldsymbol{\Delta} \end{aligned} \tag{5}$$

Next, under the event $\{\|\frac{1}{n}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^\star))\|_{\max} \le \frac{1}{2}\lambda_{lasso}\}$, we have

$$(\nabla\ell_n(\boldsymbol{\beta}^\star))'\boldsymbol{\Delta} \ge -\frac{1}{2}\lambda_{lasso}\|\boldsymbol{\Delta}\|_{\ell_1}. \tag{6}$$

Now under the same event, we combine (3), (5), (6) and the restricted eigenvalue condition (C2) to obtain

$$
\begin{aligned}
F(\mathbf{\Delta}) &> \frac{1}{3}\kappa_{logit}\|\mathbf{\Delta}\|_{\ell_2}^2 - \frac{1}{2}\lambda_{lasso}\|\mathbf{\Delta}\|_{\ell_1} + \lambda_{lasso}(\|\mathbf{\Delta}_{\mathcal{A}^c}\|_{\ell_1} - \|\mathbf{\Delta}_{\mathcal{A}}\|_{\ell_1}) \\
&\geq \frac{1}{3}\kappa_{logit}\|\mathbf{\Delta}\|_{\ell_2}^2 - \frac{3}{2}\lambda_{lasso}\|\mathbf{\Delta}_{\mathcal{A}}\|_{\ell_1} \\
&\geq \frac{1}{3}\kappa_{logit}\|\mathbf{\Delta}\|_{\ell_2}^2 - \frac{3}{2}\lambda_{lasso}\cdot s^{1/2}\|\mathbf{\Delta}\|_{\ell_2} \\
&= \frac{5s\lambda_{lasso}^2}{6\kappa_{logit}} \\
&> 0.
\end{aligned}
$$

This completes the proof of Theorem 5. □

References.

Bach, F. (2010), Self-concordant analysis for logistic regression, *Electronic Journal of Statistics*, 4, 384–414.

Fan, J., Xue, L. and Zou, H. (2013). Stong Oracle Optimality of Folded Concave Penalized Estimation.

Negahban, S., Ravikumar, P., Wainwright, M. and Yu, B. (2012), A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, *Statistical Science*, 27, 538–557.

Jianqing Fan and Lingzhou Xue
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, New Jersey 08544
USA
E-mail: jqfan@princeton.edu
lzxue@princeton.edu

Hui Zou
School of Statistics
University of Minnesota
Minneapolis, Minnesota 55455
USA
E-mail: zouxx019@umn.edu